AI 보안 활용 사례

생성형 AI와 규칙 기반 분석을 결합한 의심메일 탐지 AI Agent

메일 목록에서 위험 가능성을 사전에 식별하고, 명확한 패턴은 규칙으로 빠르게 처리하며 애매한 경우만 온디바이스 생성형 AI가 보완하는 하이브리드 보안 Agent



핵심 방향: AI가 잘하는 일과 규칙이 잘하는 일을 분리하고, 사용자 PC에서 동작하는 온디바이스 AI를 활용해 실사용 가능한 보안 Agent로 구현

01

어떤 상황에서 AI를 활용했나요?

현장의 문제 - AI Agent 필요성 - 정책적 도입 필요성

1. 기존 현장의 문제점

업무 환경에는 정상 업무메일과 함께 외부기관 안내, 광고, 세미나 초청, 결제·계정 알림을 위장한 피싱·사칭메일이 함께 유입될 수 있다.

최근 피싱메일은 정상 카드사·은행·쇼핑몰·공공기관의 안내처럼 자연스럽게 작성되어, 사용자가 메일을 열어보기 전에는 위험 여부를 구분하기 어렵다.

2. AI Agent의 필요성

사용자가 매번 발신자·제목·브랜드명·표현 방식을 직접 판단하는 대신, 메일이 화면에 나타나는 순간 자동 분석하고 위험 가능성을 알려주는 보조 Agent가 필요했다.

키워드만 쓰면 오탐이 많고, AI만 쓰면 미세한 브랜드 변조를 놓치거나 모든 메일을 분석해야 해 비효율적이었다.

3. 정책적 도입 필요성

공공·연구기관에서 생성형 AI를 활용할 때는 업무정보의 외부 전송 가능성을 함께 고려해야 한다.

본 과제는 Chrome의 온디바이스 Prompt API를 활용해 분석을 사용자 PC에서 수행하는 방향을 택했다. 민감정보 외부전송 부담을 낮추면서 생성형 AI를 업무에 적용할 수 있는 방식이다.

02

어떤 AI를 어떻게 활용했나요?

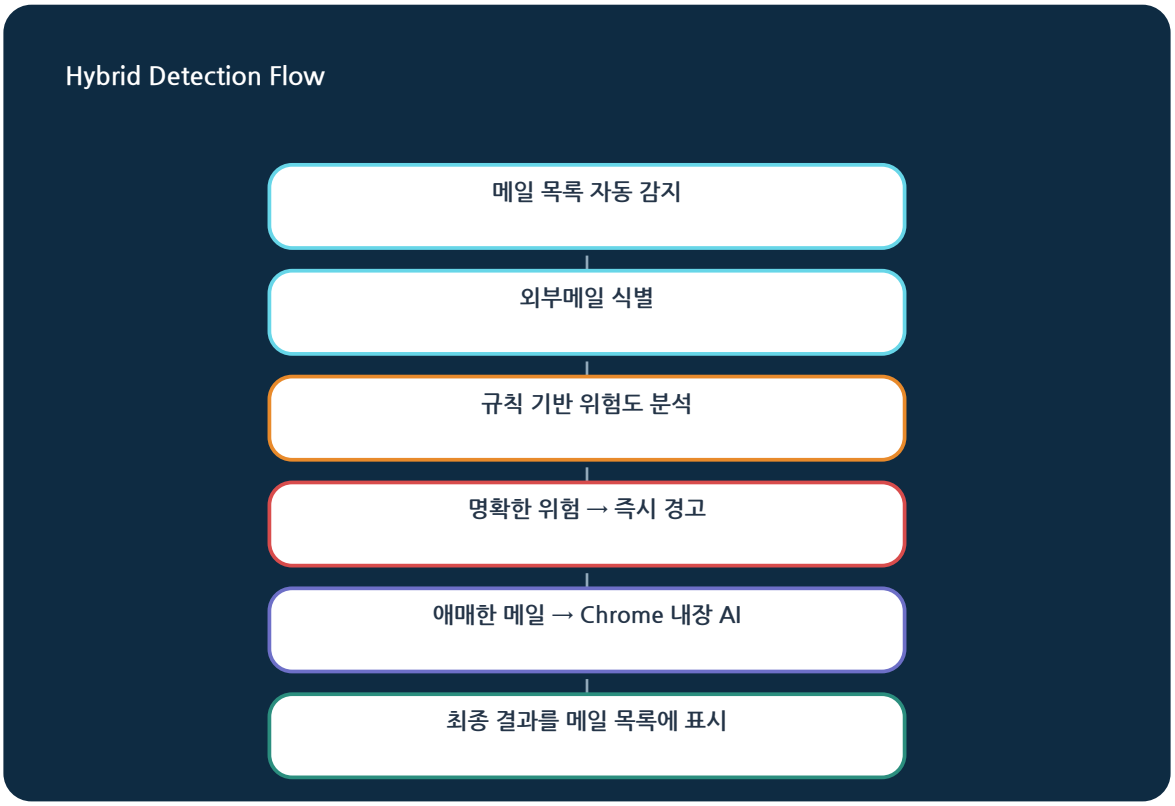
개발 단계의 AI 협업 + 실제 서비스의 온디바이스 AI

개발 과정: ChatGPT 활용

DOM 구조 분석, Chrome Extension Manifest V3 구성, JavaScript 작성, MutationObserver 기반 동적 화면 감지, Prompt API 연동, 오류 분석과 반복 개선에 활용했다.

실제 서비스: Chrome 내장 AI

Chrome Prompt API의 LanguageModel을 이용해 외부메일의 발신자와 실제 제목을 분석한다. 규칙만으로 판단하기 어려운 메일을 AI가 추가 분류한다.



03

규칙 기반 + 생성형 AI, 역할을 분리하다

빠른 1차 탐지와 문맥 기반 추가 분석의 결합

규칙 기반 1차 탐지

패턴이 명확한 경우 AI까지 호출하지 않고 즉시 판단한다.

- 브랜드명 변조
- 해외 결제·승인
- 계정 정지·잠금
- 비밀번호 변경·재설정
- 긴급 로그인·인증
- 본인 미사용 결제
- 결제내역 확인
- 배송지 오류
- 과태료·미납 안내

Chrome 내장 AI 보완 분석

규칙만으로 판단하기 어려운 메일은 여러 건을 Batch로 묶어 AI가 추가 판단한다.

- 계정 탈취 가능성
- 기업·기관 사칭
- 로그인·인증 유도
- 금융사기 가능성
- 악성 스팸 가능성

AI 출력 예시

```
{
 "is_suspicious": true/false,
 "reason": "판단 사유"
}
```

04

활용 결과 어떤 변화가 있었나요?

사용자가 메일을 열기 전에 위험 가능성을 인지

업무 흐름 변화

기존

- 메일 확인
- 발신자 확인
- 제목 확인
- 본문 확인
- 사용자가 위험 여부 판단

개선 후

- 메일 목록 진입
- Agent 자동 분석
- 의심메일 즉시 시각적 표시
- 사용자가 주의하여 확인

사용자 화면 예시

정상 교육안내	세미나 참석권(선불권) 사전구입 관련 혜택 안내	
외부메일 / 브랜드 사칭	**[현대카드] 해외 결제 승인 내역 확인 요청**	위험
내부 업무메일	주간보고 회의록	

05

나만의 방식 또는 개선 포인트

실제 테스트에서 발견한 문제를 반복적으로 개선

01

AI를 무조건 호출하지 않음

명확한 패턴은 규칙으로 즉시 처리하고 애매한 메일만 AI에 전달하여 분석량과 지연을 줄였다.

02

브랜드 사칭 오탐 조정

ST≈SK, OTI≈OCI, SS≈SSG 같은 오탐을 제거했다. 제목에서만, 4글자 이상, 동일 글자수·1글자 치환일 때만 유사 브랜드로 판단한다.

03

실제 메일 제목만 추출

[개인], [업무 (직접)], 날짜·시간 같은 메타정보와 본문 미리보기를 제외하고 실제 제목 후보만 분석한다.

04

동적 웹 UI 변화 대응

고정 XPath 의존을 줄이고 체크박스와 반복되는 mail row 구조를 활용하는 Adaptive DOM 탐색을 적용했다.

05

AI 모델 미준비 시에도 지속

Gemini Nano가 다운로드되지 않았거나 호출에 실패해도 규칙 기반 분석은 계속 동작하도록 fallback 구조를 적용했다.

06

다른 사람도 따라 할 수 있나요?

별도 서버 없이 Chrome Extension 형태로 구성

manifest.json

Chrome Extension 설정

적용 도메인, background service worker, popup 등 확장 프로그램의 기본 구성을 정의

content.js

메일 화면 감시·시각화

메일 행 탐색, 발신자/제목 추출, MutationObserver, 위험 메일 표시

background.js

탐지 엔진

규칙 기반 점수 계산, Prompt API 호출, Batch 처리, 결과 반환

popup.html / popup.js

AI 모델 준비 UI

Gemini Nano 상태 확인, 모델 다운로드 진행률, 연결 테스트

재사용 가능한 구조

메일 UI Adapter + 규칙 기반 탐지 Engine + 온디바이스 AI Engine + 사용자 경고 UI

웹메일별 DOM을 읽는 Adapter만 바꾸면 다른 메일 서비스에도 동일한 탐지 엔진을 재사용할 수 있고, 위험 키워드·브랜드 목록도 기관별 환경에 맞춰 쉽게 확장할 수 있다.

CONCLUSION

“생성형 AI를 붙이는 것”보다
AI가 잘하는 일과 규칙이 잘하는 일을 분리했다.

현장 문제 해결	메일을 열기 전 목록 단계에서 위험 가능성을 사전 인지
하이브리드 설계	규칙의 일관성 + 생성형 AI의 문맥 이해를 결합
배포 가능성	온디바이스 AI, 모델 준비 UI, fallback, Adaptive DOM까지 구현

향후 방향

발신 도메인과 링크 도메인 비교, 첨부파일 정보 연계, 실제 피싱메일 데이터셋을 활용한 정량 평가를 추가하면 탐지 정확도와 실사용 신뢰도를 더 높일 수 있다.